Detecting Unintended Social Bias in Toxic Language Datasets

Anonymous ACL submission

Abstract

Disclaimer: *This paper contains material that many will find offensive or hateful.*

Hate speech and offensive texts are examples of damaging online content that target or promote hatred towards a group or individual member based on their actual or perceived features of identification, such as race, religion, or sexual orientation. Sharing violent and offensive content has had a significant negative impact on society. These hate speech and offensive content generally contains societal biases in them. With the rise of online hate speech, automatic detection of such biases as a natural language processing task is getting popular. However, not much research has been done to detect unintended social bias from toxic language datasets. In this paper, we introduce a new dataset from an existing toxic language dataset, to detect social biases along with their categories and targeted groups. We then report baseline performances of both classification and generation tasks on our curated dataset using transformer-based models. Our study motivates a systematic extraction of social bias data from toxic language data.

1 Introduction

It is easier than ever to freely express thoughts on a wide range of topics in the age of social media and communications. This openness leads to a flood of beneficial information that can help people be more productive and make better decisions. According to a research, the global number of active social media users has just surpassed four billion, accounting for more than half of the world's population. During the next five years, the user base is predicted to continuously increase. Various studies(Plaisime et al., 2020) says that children and teenagers, who are susceptible, make up a big share of social media users. Unfortunately, this increasing number of social media users also leads to increase in toxicity(Matamoros-Fernández and Farkas, 2021). Sometimes these toxicity give birth to violence and hate crimes. It not just affect an individual, most of the time whole community suffer from its severity.

The movies and television shows we watch, and the books and articles we read, as well as the social media and meetings in which we participate and the people we surround ourselves with, all influence us. We have different perspectives based on our race, gender, religion, sexual orientation, and a whole array of other factors. These perspectives sometimes lead to biases that influence how we see the world, even if we are not conscious of them. Biases like this have the potential to lead us to make decisions that are neither intelligent nor just. And when these biases are expressed in the form of hate speech and offensive texts, it becomes painful for certain community. While some of these biases are implied, most of the explicit biases can be found in the form of hate speech and offensive texts.

More generally, in this paper, we expand on the above ideas by proposing a novel multi-level hierarchical annotation schema that encompasses the following two general categories:

A: Detection of Bias

B: Categorization of Bias and its targeted group

In the following section we discuss various established works which are aligned with our work. In section 3, we discuss about the dataset creation process which is followed by experiments and evaluations in section 4.

2 Related Work

Unfortunately, offensive content poses some unique challenges to researchers and practitioners. First and foremost, even defining what qualifies as abuse/offensive is not straightforward. Unlike other types of malicious activity, e.g., spam or malware, the accounts carrying out this type of behavior are usually controlled by humans, not bots(Founta et al., 2018).The term "offensive lan-

guage" describes a broad category of content that includes hate speech, profanity, threats, cyberbully and various ethnic and racial slurs (Kaur et al., 2021).

Hate Speech is a speech that targets disadvantaged social groups in a manner that is potentially harmful to them(Davidson et al., 2017). According to (Fortuna and Nunes, 2018), Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used.

While a lot of the research has been done to detect these toxic languages, the biased association of different marginalized groups is still a major challenge in the models trained for toxic language detection(Kim et al., 2020; Xia et al., 2020). This is mainly due to the bias in annotated data which creates the wrong associations of many lexical features with specific labels(Dixon et al., 2018). Lack of social context of the post creator also affect the annotation process leading to bias against certain communities in the dataset(Sap et al., 2019). In (Davidson et al., 2019) only racial bias in five different sets of Twitter data annotated for hate speech and abusive language are studied. A similar study have been done in (Sap et al., 2020), where dataset has both categorical and free-text annotation and generation framework as core model.

3 Dataset

People frequently hold prejudices, stereotypes, and discrimination against those outside their own social group. Positive and negative social bias refers to a preference for or against persons or groups based on their social identities (e.g., race, gender, etc.). When individuals act on their biases, this is considered discrimination. Members of certain social groups (e.g., Women, Muslims, Transgender people) are more likely to experience discrimination due to living in a society that does not adequately promote equity. In this study, we have considered five prevalent social biases:

- **Gender:** Favoritism towards one gender over other. It can be of the following types: Alpha, Beta or Sexism(Park et al., 2018).

- **Religion:** Bias against individuals on the basis of religion or religious belief. e.g. Christianity, Islam, Scientology etc(Muralidhar, 2021).

- **Race:** Favouritism for a group of people having common visible physical traits, common origins, language etc. It is related to dialect, color, appearance, regional or societal perception(Sap et al., 2019).

- **LGBTQ:** Prejudice towards LGBTQ community people. It can be due to societal perception or physical appearance.

- **Political:** Prejudice against/towards individuals on the basis of their political beliefs. For example: liberals, conservatives, etc.

Categories	Targets
Political	liberal, conservative, feminist, etc.
Religion	christian, jew, hindu, atheist, etc.
Gender	men, women
LGBTQ	gay,lesbian,homosexual,etc.
Race	black, white, asian, canadians,etc.

Table 1: Bias annotation categories and targets

.

For all of these categories, target terms are the communities towards which bias is targeted.

3.1 Annotation

The dataset we used for annotation is collected from a Kaggle competition named "Jigsaw Unintended Bias in Toxicity Classification (jig, 2019; Research Data, 2018)". It has around 2M Wikipedia comments annotated for toxicity. The data also has several other toxicity subtype attributes such as severe toxicity, obscene, threat, insult, identity_attack, sexual_explicit.

On observing this dataset closely, we found that all of the columns are redundant for bias detection task except the identity_attack column. Identity_attack shows denigration of the comment towards some identity. This identity could be related to race, religion, political, gender or sexual orientation. So, we filtered out only those comments for which identity_attack values are greater than or equal to 0.5. On observing this filtered dataset closely, we found that many of them contain some kind of social bias. We have considered only five bias categories for our annotation, which are listed in the Table 1 along with targets. We did not include other categories due to their low presence in

comment text	label	category	target	implication
I wish gay men would just leave children alone.	biased	lgbtq	gays	gays are pedophiles
Islam, unfortunately, is a violent cult.	biased	religion	muslims	muslims are violent
Why does Netflix hate gay people?	unbiased	-	-	-
Liberal women are so #f̂king ugly!	biased	political,gender	liberals,women	hatred towards liberal women

Table 2: Annotation of Bias along with its category, targeted groups and implication

the original dataset. Along with types and targets, implications are also annotated. Sample annotation of this filtered dataset is shown in Table 2.

A total of 3000 instances were annotated with multiple labels(Refer A.3 and table 6 for more details). To check the consistency of our framework and to categorize biases, two different annotators annotated the data independently. Inter-annotator agreement for first 1500 instances were calculated and the Cohen's Kappa (Hsu and Field, 2003) score of 50.3 was observed with total agreement score to be 86%. The distribution of data among multiple categories is shown in the figure 1. All the disagreement between annotators were resolved by adjudication with the help of an expert.

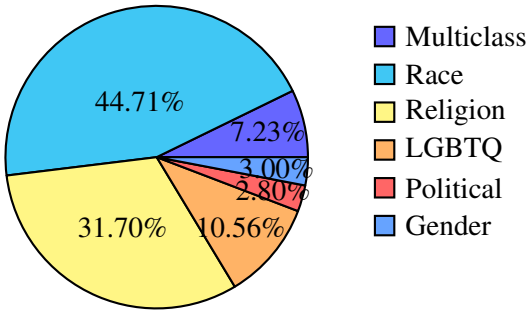

Figure 1: Distribution of data among different categories. It is observed that some instances qualified for multiple bias categories(7.23%).

3.2 Annotation Challenges

While annotating toxic datasets, several challenges have been observed, which needs to be resolved in order to annotate and create a consistent dataset. First, is quoting someone else statement considered as bias? For example:
Trump said "Mexicans are rapists and drug dealers".

While one may argue that such statements are not biased as they are simply quoting someone else opinions that are not theirs, but we decided to annotate them as biased because quoting someone's else statement is equally harmful and damaging.

On the other hand, we believe that asking questions about an issue may not lead to bias. For example:
Black idiot or white idiot. What is the difference?

For this statement, two schools of thought may emerge. One could think that this statement is about an idiot in general without discriminating on the basis of race. So it will not be a bias. On the other hand, for some people, this statement could also mean that both the blacks and whites are referred as idiots here.

We also encountered statements lacking context and statements made as a personal attack. These instances were not flagged as bias. Some sarcastic instances were also observed and were labelled appropriately.

4 Experiments and Evaluation

In this section we will be discussing about different models trained for detection of social biases and their categories. For all our experiment, we split the data into train(0.75), dev.(0.10), test(0.15). Before bias detection, we started with identity attack detection, which is a trivial binary class classification. For this we used a BI-LSTM(Huang et al., 2015) with two word embeddings, FastText crawl 300d (Bojanowski et al., 2017) and Glove 300d(Pennington et al., 2014). After identity attack detection, several model architectures were tried to detect biases, which is our primary goal. We will discuss each of them in detail in the following subsections.

4.1 Hierarchical Model

In hierarchical model, identity attack detection, bias detection and category classification was done in three level respectively. The identity attack was detected using above LSTM model. 12 layer BERT base uncased was used to detect biases and their categories subsequently.

4.2 Multi-task Learning

Multi Task Learning, in the context of classification, tries to improve the performance of numerous classification problems by learning them together.

Model	P	R	F1	Acc
Hierarchical	**0.84**	0.56	0.57±0.01	**0.84**
Multi-task	0.41	0.50	0.45±0.00	0.82
GPT2	0.72	**0.61**	**0.63±0.06**	0.81

Table 3: Results for bias detection. Standard deviation of 3 runs(for F1) are reported.

Model	P	R	F1	Acc
Hierarchical	**0.81**	0.75	0.76±0.02	0.93
Multi-task	0.77	0.75	**0.77±0.02**	0.93

Table 4: Results for Bias Category Detection

.

So instead of predicting bias and its category in two steps, we can train a model to predict them simultaneously in one step. This model was also built using BERT base uncased (Devlin et al., 2019). It had to predict two labels. In label 1 it would predict whether the text has bias or not and in label 2 it will predict the bias categories such as race, religion, LGBTQ, political, and/or gender.

We have reported precision(P), recall(R), F1 (macro values for all), and accuracy(Acc) for both experiments with best numbers in bold.

4.3 Generation Framework

Considering the efficacy of GPT(Radford and Narasimhan, 2018) based model for classification, conditional generation tasks(Sap et al., 2020), we frame the prediction of categorical variables and implications as generation task. The input is a sequence of tokens as in Equation1, where w_i are the tokens corresponding to comment text and [BOS], [SEP], [EOS] are start token, separator token and end token respectively. Two task specific tokens([BON], [BOFF]) were added to the token vocabulary which were used as $w_{[\text{bias }]}$ in the input. As we have many inputs with multiple bias categories and targets, we combine them using a comma separator in the raw text. While encoding the input we use $w_{[\text{C}]_i}, w_{[\text{T}]_i}$ as the token corresponding to them respectively. Similarly, $w_{[\text{R}]_i}$ is used for representing the tokens corresponding to implications.

$$\mathbf{x} = \{[\text{BOS}], w_i, [\text{SEP}]\ w_{[\text{bias }]}, [\text{SEP}] \\ w_{[\text{C}]_i}, [\text{SEP}] w_{[\text{T}]_i}, [\text{SEP}] w_{[\text{R}]_i}, [\text{EOS}]\} \quad (1)$$

For this experiment, we finetune the GPT-2 (Radford et al., 2018) model with commonly used hyperparameters. For training we use cross-entropy loss as cost function. During inference, we first calculate the normalized probability of $w_{[\text{bias }]}$ conditioned on the initial part of input and then append the highest probable token to the input and generate rest of the tokens till [EOS].

We use BLEU-2 (Papineni et al., 2002) and RougeL(Fmeasure) (Lin, 2004) as the metrics to calculate the performance of the model for category, target and implication of the comment text(Table 5) and macro F1 as metric for bias evaluation(Table 3). Performance for category generation is better than other two variable as it has less ambiguity whereas the low performance for implications show the variability in the annotation for implications.

Variables	BLEU-2	RougeL
Categories	61.60±0.96	88.23±1.23
Target names	52.95±2.84	77.58±4.21
Implications	33.4±1.55	39.5±1.20

Table 5: Evaluation of various generation tasks. The standard deviations for 3 runs are also reported.

5 Conclusion and Future Work

We have shown that identity attacks or hate speech generally contain some kind of social bias or stereotypes in them. However not all hate speech can be labelled as biased. Some of them are merely personal attacks. We observed that many times detecting bias without context for the comment or demographics information of the comment holder makes the annotation much more challenging. Filtering out such biases from hate speech is not a trivial task. Our best model could only get an F1 score of 0.63. However, detecting these social bias from toxic datasets, which are available in relatively large amount, will be useful starting point for social bias research in other forms of text.

At the time of inference, the problem of model bias was also encountered. Merely the presence of certain community words (Muslim, whites, etc.) make model to label a comment as social bias. This indicates that more sophisticated models along with explainability tools are required to detect biases. In the future, we would also like to expand the annotation process and bias detection to more categories along with implicit biases which are hard to be detected by an AI model.

References

2019. Jigsaw unintended bias in toxicity classification.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51:1–30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Louis M. Hsu and Ronald Field. 2003. Interrater agreement measures: Comments on kappan, cohen's kappa, scott's , and aickin's . *Understanding Statistics*, 2(3):205–219.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281. AI in Computational Linguistics.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

Deepa Muralidhar. 2021. *Examining Religion Bias in AI Text Generators*, page 273–274. Association for Computing Machinery, New York, NY, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marie Plaisime, Candace Robertson-James, Lidyvez Mejia, Ana Núñez, Judith Wolf, and Serita Reels. 2020. Social media and teens: A needs assessment exploring the potential role of social media in promoting health. *Social Media + Society*, 6(1):2056305119886025.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Civil Research Data. 2018. Civil comments.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language.

Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

A Appendix

A.1 Ethical Considerations

Our work aims at capturing various social biases in toxic social media posts and demonstrates the annotation quality on biases in one of existing dataset. We also discuss the challenges we faced while doing the annotation of the dataset, specifically due to the absence of context for each instance in the dataset. Also, study of social biases come with ethical concerns of risks in deployment (Ullmann and Tomalin, 2020). As these toxic posts can create potentially harm to any user or community, it is required to conduct this kind of research to detect them. If done with precautions, such research can be quite helpful in automatic flagging of toxic and harmful online contents.

Researchers working the problem of social bias detection on any form of text would benefit from the dataset we have collated and from the inferences we got from multiple training strategies. **Limitations** Our work currently considers only five types of social biases; not all other possible dimensions of bias. As discussed in section 5, our model sometimes predicts a comment as biased which has mentions of some community words but there is no potential harm in it.

A.2 Annotator Demographics and Treatment

Both the annotators were trained and selected through extensive one-on-one discussions, and were working voluntarily. Both of them went through few days of initial training where they would annotate many examples which would then be validated by an expert and were communicated properly about any wrong annotations during training. As there are potential negative side effects of annotating such toxic comments, we used to have regular discussion sessions with them to make sure they are not excessively exposed to the harmful contents. Both the annotators were Asian male and were of age between 23 to 26. The expert was an Asian female with post-graduation degree in sociology.

Categories	train	dev	test
bias	1848	246	371
neutral	401	53	81
race	921	117	188
religion	656	99	126
gender	95	11	19
lgbtq	237	26	46
political	98	10	16

Table 6: Distribution of different categories acros 3 splits of Train, Dev. and Test.

A.3 More Details about Dataset

In the original dataset, the identity_attack label show aggregated rating given by multiple annotators if there is presence of hate towards any identity group. So we decided to curate the dataset by conditioning on the identity_attack label. We curated 3000 instances for bias detection with five possible bias categories. Data was annotated with multiple labels which have 120 unique words for target annotation across five categories. All categories and corresponding target names are mentioned in table 1.

Out of 3000, out dataset has 2465 has bias annotations (82% of dataset) and 535 are neutral(not biased towards any identity). The number of instances for each category across train, dev., test are shown in table 6.

A.4 Training Details

A.4.1 BERT Training

We finetune 12 layer BERT base uncased with batch size of 32 for two epochs. Max token length of 64 is used. We also use a dropout layer in our model. Adam optimizer with learning rate = 5e-05, epsilon = 1e-08, decay = 0.01, clipnorm = 1.0 were used.

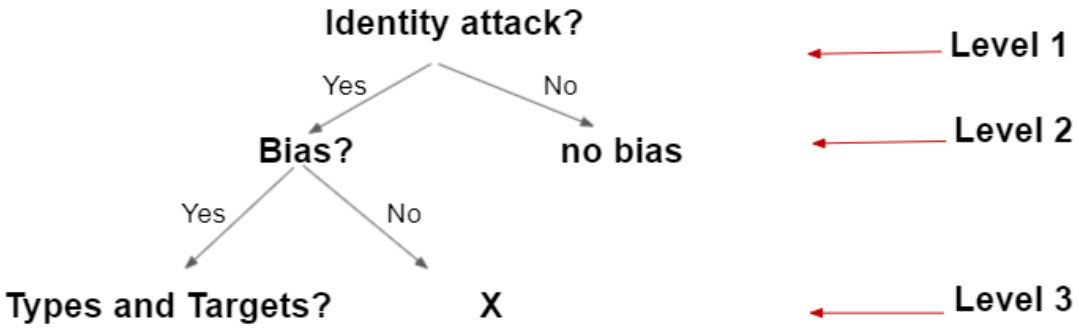

Figure 2: Hierarchical structure to annotate Bias and its Categories

A.4.2 GPT-2 Training

We finetune GPT-2 with a training batch size of 1, gradient accumulation step as 4, and 200 warm up steps. We experiment with learning rates of 2e-5,3e-5, 4e-5, 5e-5 with AdamW(Loshchilov and Hutter, 2019) optimizer and epochs of 5, 10, 20. Experiments were run with a single GeForce RTX 2080 Ti GPU. Finetuning one GPT-2 model took around 40 minutes for 5 epochs.

All of our implementations uses Huggingface's transformer library(Wolf et al., 2020).

A.5 Annotation Guidelines

In the Jigsaw Unintended Bias in Toxicity Classification Data, a comment is labelled for toxicity and its various sub-types such as severe_toxicity, obscene, threat, insult, identity_attack, sexual_explicit. All the values are between 0 and 1 which represent the fraction of human raters who believed the attribute applied to the given comment.

We followed a three level hierarchical annotation process to annotate this data for bias. In Level 1, we filtered out all those instances for which identity attack values are greater than or equal to 0.5. We choose identity attack as our filtering criteria because when someone attacks someones identity(race,religion,gender, etc.) then he/she is probably showing prejudice towards a community on the basis of its identity and therefore these attacks are highly biased.

On further analysis, we found that not all identity attacks are biased. Some of them were just personal attacks and few of them lacked context to be marked as biased . So, in Level 2, annotators were instructed to mark 1 if the comment has bias otherwise 0. Before this level, all the annotators were given definition of bias, i.e.

Bias: Bias refers to being in favour or against/ preference or prejudice towards certain individuals, groups or communities based on their social identity (i.e., race, gender, religion etc.). Bias can be

- **positive or negative**
- **based on stereotypes:** It is an overgeneralized belief about a particular section of population or community. For example: "Asians are good in maths" but other people are also good in maths.
- **Bias is an individual preference:** For example: if you hire an Asian for a job that also has an equally qualified black applicant because you think blacks are not as smart as Asians, then this is bias.

If the comment was found to be biased in Level 2, then bias types and targets along with implied meaning were annotated in Level 3. Otherwise no further annotation were performed in Level 3. Bias types and targets corresponding to each types is already given in Table1. The whole three level hierarchical scheme can be understood clearly by the decision tree diagram shown in Figure 2